Exploring Consumer Exposure Pathways and Patterns of Use for Chemicals in the Environment

Kathie L. Dionisio[a] (dionisio.kathie@epa.gov), Alicia M. Frame[b,1] (amframe@dow.com), Michael Rock Goldsmith[a,2] (rocky.goldsmith@gmail.com), John F. Wambaugh[b] (wambaugh.john@epa.gov), Alan Liddell[c,3] (acliddel@ncsu.edu), Tommy Cathey[d] (cathey.tommy@epa.gov), Doris Smith[b] (smith.doris@epa.gov), James Vail[b] (vail.james@epa.gov), Alexi S. Ernstoff[e] (alexiernstoff@gmail.com), Peter Fantke[e] (pefan@dtu.dk), Olivier Jolliet[f] (ojolliet@umich.edu), Richard S. Judson[b,*] (judson.richard@epa.gov)

[a]U.S. Environmental Protection Agency, National Exposure Research Laboratory, 109 T.W. Alexander Drive, MC E205-02, Research Triangle Park, NC, 27709, USA

[b]U.S. Environmental Protection Agency, National Center for Computational Toxicology, MC B205-01, Research Triangle Park, NC, 27709, USA

[c]North Carolina State University, Department of Mathematics, Box 8205, Raleigh, NC, 27695-8205, USA

[d]Lockheed Martin, Research Triangle Park, NC, USA

[e]Quantitative Sustainability Assessment, Department of Management Engineering, Technical University of Denmark, Produktionstorvet 424, 2800 Kgs. Lyngby, 2800, Denmark

[f]Department of Environmental Health Sciences, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI, 48109, USA

Affiliation listed above is where work was completed. Current affiliations listed below where applicable.

[1]U.S. Environmental Protection Agency, Office of Solid Waste and Emergency Response, Arlington, VA 22202, USA

[2]Chemical Computing Group, Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, Quebec, H3A 2R7, Canada

[3]North Carolina State University, Department of Mathematics, Box 8205, Raleigh, NC, 27695-8205, USA

*Corresponding author:
Richard Judson
U.S. Environmental Protection Agency
109 T.W. Alexander Drive (B205-01)
Research Triangle Park, NC 27709
phone: 919.541.3085
fax: 919.541-1194
judson.richard@epa.gov

Highlights:

- To assign use-related information to chemicals to help prioritize which will be given more scrutiny relative to human exposure potential
- Categorical chemical use and functional information are presented through the Chemical/Product Categories Database (CPCat)
- CPCat contains information on >43,000 unique chemicals mapped to ~800 terms categorizing their usage or function
- The CPCat database is useful for modeling and prioritizing human chemical exposures

Abbreviations:

ACToR: Aggregated Computational Toxicology Resource

AICS: Australian Inventory of Chemical Substances

CAS RN: Chemical Abstracts Service Registry Number

CDR: Chemical Data Reporting Rule

CPCat: Chemical/Product Categories Database

DCPS: Danish Consumer Product Survey

DfE: Design for the Environment

EDSP: Endocrine Disruptor Screening Program

EPA: Environmental Protection Agency

EWG: Environmental Working Group

GRAS: Generally Recognized as Safe

HTP: Human Toxome Project

IUR: Inventory Update Reporting Modifications Rule

MSDS: Material Safety Data Sheets

NICNAS: National Industrial Chemicals Notification and Assessment Scheme

RPC: Retail Product Categories Database

SDWA: Safe Drinking Water Act

SPIN: Substances in Preparation in Nordic Countries

TSCA: Toxic Substances Control Act

Disclaimer:

Although this work was reviewed by EPA and approved for publication, it may not necessarily reflect official Agency policy. EPA does not endorse the purchase of any commercial products or services mentioned in this publication. The authors declare they have no actual or potential competing financial interests.

Abstract

Humans are exposed to thousands of chemicals in the workplace, home, and via air, water, food, and soil. A major challenge in estimating chemical exposures is to understand which chemicals are present in these media and microenvironments. Here we describe the Chemical/Product Categories Database (CPCat), a new, publically available (http://actor.epa.gov/cpcat) database of information on chemicals mapped to "use categories" describing the usage or function of the chemical. CPCat was created by combining multiple and diverse sources of data on consumer- and industrial-process based chemical uses from regulatory agencies, manufacturers, and retailers in various countries. The database uses a controlled vocabulary of 833 terms and a novel nomenclature to capture and streamline descriptors of chemical use for 43,596 chemicals from the various sources. Examples of potential applications of CPCat are provided, including identifying chemicals to which children may be exposed and to support prioritization of chemicals for toxicity screening. CPCat is expected to be a valuable resource for regulators, risk assessors, and exposure scientists to identify potential sources of human exposures and exposure pathways, particularly for use in high-throughput chemical exposure assessment.

Keywords: chemical exposure, human exposure, high throughput, exposure prioritization, use category

1. Introduction

As high-throughput hazard screening approaches such as ToxCast and Tox21 (Dix et al. 2007; Judson et al. 2010; Kavlock and Dix 2010; Kavlock et al. 2012; Tice et al. 2013) continue to evolve, there is a need to develop methods to obtain high-throughput exposure estimates so that chemical hazard screening approaches and exposure estimates together can allow for the complete development of high-throughput risk models. A major challenge in estimating the risk of chemical exposures to human health is the lack of consistent information describing how chemicals are used. A limited number of chemicals that are known to have biological targets, and with uses that suggest high exposures such as pharmaceuticals and some food use pesticides, have been well characterized on both the hazard and exposure axes. For the remaining majority of marketed chemicals, there is little publicly available information (Egeghy et al. 2012; Judson et al. 2009). This information is critical since the presence of a chemical in specific products significantly influences the nature and extent of human exposures. While information on production volume of chemicals is currently available, a large, uniformly organized repository of information on how chemicals are used, product composition, and other properties (e.g. physicochemical form of the chemical within the product) currently does not exist. This paper describes an effort to characterize one component of the high-throughput exposure estimation process: categorizing the usage of chemicals.

To address the deficiency in chemical exposure estimates, previous efforts have utilized relatively simple high-throughput environmental and indoor fate and transport models that have been parameterized using widely available molecular descriptors, such as physicochemical properties, and simple binary descriptors of indoor and consumer use (Mitchell et al. 2013; Wambaugh et al. 2013). Specifically, researchers from our group have shown that the simple metric of presence or

absence of a chemical in consumer products and associated indoor use is an indicator of a chemical being above the limit of detection for biomonitoring (Wambaugh et al. 2013). Although useful for specific applications, the uncertainty bounds on these models are relatively large and additional information on product and chemical use would help to refine these models. To fill this gap in knowledge, we present here the Chemicals/Products Categories Database (CPCat), the result of a large-scale effort to catalog and consolidate relatively disparate data sources in order to make chemical use information publicly available, and in a form useful for high-throughput exposure modeling. This new database provides critical information for comparing between well-studied and novel chemicals with respect to use – a key factor driving human exposure to these chemicals. Aggregating publicly available data sources which categorize chemicals using terms describing their usage, and merging these diverse sources into a single data set with consistent chemical use categories, is the first step towards integrating chemical use data into high-throughput exposure models. We have compiled an extensive list of chemicals and their associated categories of chemical and product use. Unique use category taxonomies from each source are mapped onto a single common set of terms. We provide several examples of the application of the database that identify and enumerate chemical exposure pathways, including: (1) identifying all documented potential uses of a specific chemical; (2) cataloging all chemicals that meet an exposure scenario (e.g., exposure from children's products); and (3) examining the potential uses of the chemicals implicated with a specific adverse outcome pathway (AOP) (Ankley et al. 2010) (e.g. for use in the U.S. Environmental Protection Agency (EPA) Endocrine Disruptor Screening Program (EDSP)). We anticipate that this open-source database will grow as relevant data continues to become available and is integrated into CPCat, and that this resource will be useful in chemical exposure research and to regulatory agencies.

144 ## 2. Methods

145 Here we describe the methodology used to construct the relational CPCat database, available for
146 public download, and as an online searchable website, at http://actor.epa.gov/cpcat. Our approach
147 to developing CPCat involved collecting a variety of publicly available data on chemicals and
148 associated categorical (use-categorization) groupings, annotating and curating these data, and
149 harmonizing these categories into a single set of terms. CPCat integrates information from major
150 national and international sources to provide categorical groupings for 43,596 unique chemicals.

151 ### 2.1 Classes of chemical use categories

152 Chemical use categories as defined by the data sources can be grouped into 5 general classes.
153 When a chemical has a variety of documented uses and functions, it may be associated with
154 multiple classes, and/or multiple categories within each class (Table 1).

155

156 **Table 1:** Classes of chemical use categories

Class	Definition
General-use	General categories for chemicals which do not fall into any of the more specific classes of chemical use categories defined below (e.g., lipstick)
Product-use	Categories taken from classifications used for retail products (e.g., children's toys)
Therapeutic-use	The chemical is used as an ingredient in a pharmaceutical, with categories defined by the type of ailment being treated (e.g., anti-acne)

Functional-use	Categories defined by the chemical's properties, which determine the chemical's use; does not specify the type of product in which the chemical is performing the function (e.g., a solvent)
Industrial sector-use	The chemical is used in an industrial sector, with categories defined by the type of industry (e.g., mining)

157

158

159 2.2 Data sources

160 Multiple data sources, including information provided by companies, trade associations, and
161 regulatory agencies, were used to construct the CPCat database. Table 2 details the class of
162 chemical use category (as provided by each source), the number of specific categories (provided
163 by the source), and the number of chemicals associated with each source.

164

165 **Table 2**: Summary of data sources used to construct the CPCat database.

Original data source [a]	**Class of categories [b]**	**Original categories [c]**	**CPCat cassettes [d]**	**Chemicals**
ACToR Data Sets and Lists	General-use	131	173	35,838 [e]
ACToR UseDB	General-use	15	15	31,622
CDR 2012:				
Consumer	General-use	34	36	3,321
Industrial Function	Functional-use	34	27	5,023
Industrial Sector	Industrial sector-use	42	43	5,226

DfE	Functional-use	11	9	444
Dow	Functional-use	19	18	104
DrugBank	Therapeutic-use	582	460	1,754
2006 IUR	General-use	19	24	1,152
KemI	Functional-use	61	31	876
NICNAS	General-use	17	17	177
Retail Product Categories	Product-use	359	191	2,778
SPIN:				
detpcat	General-use	781	284	6,491
Industrial Sector	Industrial sector-use	580	221	4,603
NACE	Industrial sector-use	57	52	7,745
UC62	General-use	61	59	9,059
Toxome	Functional-use	16	16	442

166

167 [a] Source names listed match source names used in the downloadable CPCat database.

168 [b] Class of category used for chemical categorization in the original data source.

169 [c] Number of unique chemical categories in the original data source.

170 [d] The term "CPCat cassette" is defined below in section 2.3.

171 [e] Note that >550,000 chemicals are included in ACToR, but only ~36,000 could be mapped to one

172 or more use categories.

173

174

175 *Aggregated Computational Toxicology Resource (ACToR) Data Sets and Lists:* The U.S. EPA's
176 ACToR database is a compilation of publicly available data on chemical toxicity for more than
177 550,000 unique chemicals (http://actor.epa.gov) (Judson et al. 2008; Judson et al. 2009; Judson et
178 al. 2012). ACToR includes, but is not limited to, high and medium production volume industrial
179 chemicals, pesticides (active and inert ingredients), and potential ground and drinking water
180 contaminants. The ACToR database is organized around chemicals, data sets, and lists, where an
181 ACToR data set refers to data linking chemicals to physicochemical properties, bioactivity, and
182 hazard measurements and an ACToR list refers to chemicals meeting a given criteria. ACToR
183 includes many sources which were subsequently included in CPCat, through both ACToR data
184 sets and lists. Note the Danish Consumer Product Survey (DCPS;
185 h*ttp://www.mst.dk/English/Chemicals/consumers_consumer_products/danish_surveys_consumer*
186 *_products/*) is included within this source. The DCPS analyzes consumer products in laboratory
187 testing to determine if they may pose a threat by releasing chemicals to the air, or when in contact
188 with the human body. The DCPS includes information on which chemicals were detected in
189 experimental tests, and which chemicals were analyzed for but were not detected.
190 *ACToR UseDB:* The ACToR UseDB is a database of chemicals assigned to a small number of
191 broad chemical-use categories. The UseDB was created by the authors based on information
192 extracted from the ACToR database. See supplemental text for a detailed description.
193 *Design for the Environment (DfE)*: The DfE program of the U.S. EPA (www.epa.gov/dfe/)
194 evaluates human health and environmental concerns for chemicals used in a range of industries.

The program partners with various groups in order to identify safer products and ways to reduce the use of chemicals of concern. The DfE's Safer Chemical Ingredients List categorizes chemicals by functional-use (e.g. colorants, fragrances, solvents, etc.).

Dow: The Dow Chemical Company has published functional-use categorizations for many of the chemicals they manufacture, which are primarily used in the industrial sector (http://www.dow.com/productsafety/assess/finder.htm).

DrugBank: DrugBank is a database of pharmaceutical ingredients compiled by the University of Alberta, Canada, which categorizes chemicals by therapeutic-use (http://www.drugbank.ca/) (Wishart et al. 2006; Wishart 2008; Wishart et al. 2008).

U.S. EPA 2006 Inventory Update Reporting (IUR) Modifications Rule and the 2012 Chemical Data Reporting (CDR) Rule: The U.S. EPA IUR rule (now known as the CDR rule) allows the U.S. EPA to collect and publish information on the manufacturing, processing, and use of commercial substances and mixtures on the Toxic Substances Control Act (TSCA) Chemical Substance Inventory (http://cfpub.epa.gov/iursearch/). Data from both the 2006 IUR and the 2012 CDR are included here, covering primarily industrial chemicals and their corresponding use categories. Note the 2012 CDR includes three distinct data sources which categorize chemicals by general-use (for consumer products), and by functional- and industrial sector-use (for industrial chemicals).

Swedish Chemicals Agency (KemI): The Swedish KemI is a government agency responsible for ensuring the safe use of chemicals, and maintains a product registration list and variety of databases

for pesticides and other chemicals. This organization has published a list of chemicals categorized by functional-use (http://www.kemi.se/en/).

National Industrial Chemicals Notification and Assessment Scheme (NICNAS): NICNAS (http://www.nicnas.gov.au) maintains the Australian Inventory of Chemical Substances (AICS) list, a listing of industrial chemicals in use in Australia since January 1, 1977. The list categorizes chemicals by general-use, with a small number of categories.

Retail Product Categories (RPC) Database: Goldsmith et al. developed a database of chemical information extracted from publicly available Material Safety Data Sheets (MSDS) for products sold at Walmart (Goldsmith et al. 2014). In addition to extracting quantitative information on chemical composition of products from the MSDS, products and their ingredients were mapped to a hierarchy of product-use categories.

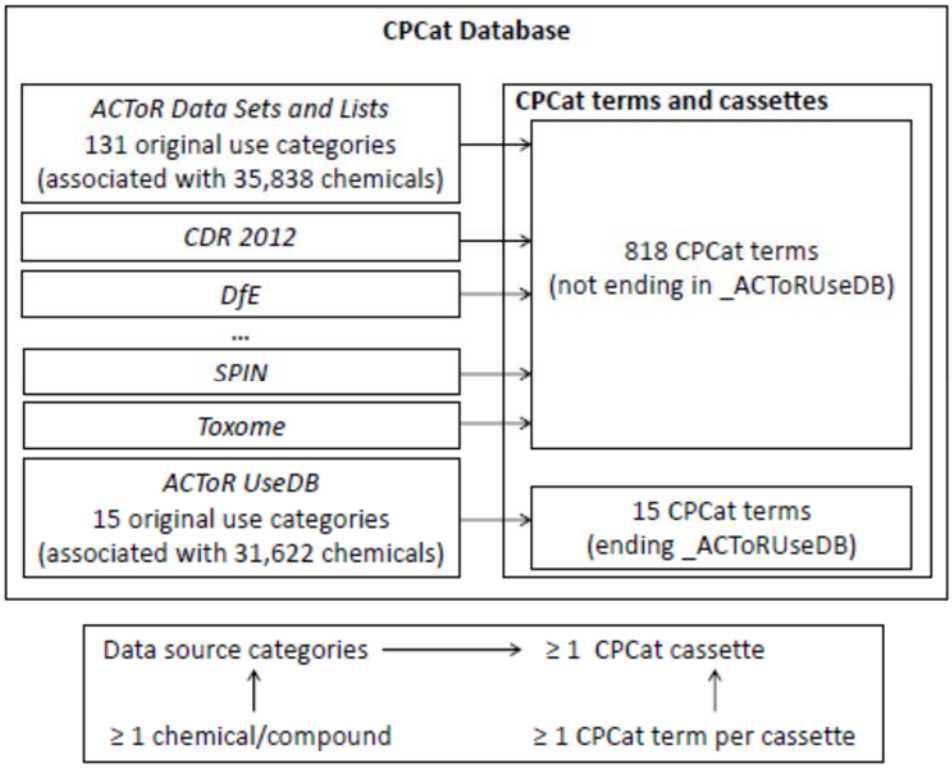
Substances in Preparation in Nordic Countries (SPIN) database: SPIN is a joint project of government environmental agencies in Norway, Sweden, Denmark, and Finland, and is comprised of data from the Product Registries of each of these countries (SPIN 2013). Four separate SPIN databases which categorize chemicals in different ways are used in constructing CPCat: old Danish and Norwegian categories (detpcat), use/function categories for chemical substances and preparations (UC62), the Statistical classification of economic activities in the European Community (NACE), and industrial-use information (Industrial Sector). The first two databases categorize chemicals by general-use, the latter two categorize chemicals by industrial sector-use.

Human Toxome Project (HTP): The Environmental Working Group (EWG) HTP collects biomarker data to help understand the scope of population-level exposure to industrial chemicals

236 that enter the body through pollution or as ingredients in consumer products
237 (http://www.ewg.org/sites/humantoxome/). Data from the HTP includes a small number of
238 categories of functional-use which have an elevated toxicity risk.

239 2.3 Assigning CPCat terms and cassettes

240 The CPCat database consists of each of chemicals for which one or more sources reported use data
241 and an associated set of CPCat terms describing usage. The terms are organized using a well-
242 defined nomenclature to create 'cassettes.' Each of the data sources used to construct the CPCat
243 database employed a unique set of chemical use categories (each falling into one of the five
244 chemical use classes described above) to meet a particular need. These tend to focus on one or a
245 few types of uses or functional categories, or on particular classes of chemicals. No single
246 categorization scheme included all of the categories covered in the global collection. To create
247 CPCat, we manually mapped the chemical use categories and descriptions provided by each data
248 source to CPCat terms and cassettes (Figure 1). Mining the use category descriptions provided
249 within each of the original data sources results in 2,681 unique original source chemical use
250 categories (noting that the same description/category can be used by more than one source), which
251 were mapped to 833 unique CPCat terms (Figure 1).



252

253 **Figure 1:** CPCat database organization

254

255 Cassettes are comprised of one or more CPCat terms, separated by spaces; all CPCat terms within
256 a cassette must be interpreted together to reflect the categorical information provided by the
257 original data source. Because of the broad nature of the 15 original ACToR UseDB categories,
258 these 15 categories were mapped directly to 15 corresponding CPCat terms (indicated by the suffix
259 "_ACToRUseDB"); no categories from other sources were mapped to these "_ACToRUseDB"
260 CPCat terms.

261 The full set of CPCat terms were selected by aggregating all categories provided by each data
262 source, taking care to eliminate synonymous category names (e.g., drug and pharmaceutical),

263 mistakes (e.g., spelling errors), and other redundancies or superfluous information. No attempt was
264 made to extrapolate or fill in missing data on chemicals. Rather, CPCat incorporates only existing
265 information on use categories for chemicals from each data source. An underscore between two
266 words indicates a compound word (e.g., automotive_care, building_material) and should be
267 considered the same as a single unique CPCat term. Any combination of CPCat terms can be
268 combined to create a CPCat cassette; however there are some combinations of terms that are
269 common, and others which never occur.

270 2.4 Interpreting CPCat terms and cassettes

271 A data dictionary including a list of all unique CPCat cassettes, and describing each unique CPCat
272 term, is included with the release of the database at http://actor.epa.gov/cpcat. While a specific
273 hierarchy was not defined, CPCat terms refer to different levels of detail, due to the varying levels
274 of information available from each source regarding the usage of the chemical. When a source
275 included specific information on the chemical usage, for example a specific type of beauty product
276 such as lipstick, that information is reflected in the assigned CPCat cassette so that information is
277 not lost. If more than one CPCat cassette was mapped to a single source category (separated by a
278 comma), this indicates that the source reported more than one distinct usage for the chemical within
279 one original category entry. In this situation, each cassette should be interpreted separately to
280 reflect these multiple uses for the chemical.

281 Examples of CPCat cassettes include: a) building_material, b) “manufacturing building_material
282 wood”, c) “building_material wood”, and d) “furniture wood”. Where a) describes a chemical with
283 a general use in building materials, but with no further information given in the original data

source; b) describes a chemical used when manufacturing wooden building materials; c) describes a chemical contained in wooden building materials; and d) describes a chemical used in wooden furniture. When a CPCat cassette is comprised of more than one term, the terms refer to increasing levels of specificity when reading from left to right. As an example of when multiple CPCat cassettes might be assigned to a single original data source category, if the original data source category described wooden furniture and housing materials, then this entry would have been assigned both c) and d) cassettes in order to reflect the multiple uses specified by the original data source entry.

Some data sources determined chemical content of a product through laboratory testing, rather than from listed ingredients. The CPCat term for this is 'detected.' Thus a chemical may appear in a use category due to unintentional inclusion of that chemical in a product (e.g., because of contamination). Any source which indicated chemicals were detected through laboratory testing (including all DCPS sources) include "detected" as a CPCat term within the associated cassette(s). Note that the quantitative data from the laboratory testing is currently not included in CPCat, rather if the presence of a chemical is detected in laboratory testing, the information is included as such in CPCat. The "child_use" and "baby_use" CPCat terms are similarly unique in that they reference the class of consumer for which the product is intended. These terms were included due to the general interest in exposure of these demographics, and due to the number of products specifically marketed to these demographics.

CPCat terms associated with the 15 broad UseDB categories (Supplemental Table 1) are unique within CPCat. Because the 15 UseDB categories are quite broad, it was desired to distinguish these

category assignments from the remainder of categorical assignments within CPCat. Then, if a user only wanted to analyze the 15 broadly defined UseDB categories and their associated chemicals, these could easily be extracted. Or, if a user wanted to exclude these broadly defined categories from their search, this could be done. The CPCat terms associated with the 15 UseDB categories include the suffix "_ACToRUseDB" to alert the user to these unique CPCat terms that indicate a potentially broad categorization of the chemical.

2.5 Data management and database availability

To aid data processing, chemical category taxonomies from each source were translated into a common format before entry into the CPCat database. For each chemical listing in the CPCat database, in addition to the assigned CPCat cassette(s), links to the underlying data source(s) and original taxonomy categories are maintained. In the database, each category is labeled by an alphanumeric ID, and a description. The top level of each source or taxonomy is always given the ID "Source_0000." When sources used an explicit ID for each category, they have been maintained in CPCat. This information is not included in the web interface.

The CPCat database is available in three formats. A .zip file containing a set of .txt and Microsoft Excel files is available for download (http://actor.epa.gov/cpcat), which includes R code for running the examples presented below. Alternatively, a MySQL database for download, and a searchable online version of the CPCat database, are available at the same location.

3. Results

3.1 Summary statistics

A total of 43,596 unique chemicals from the U.S. EPA's ACToR database mapped to at least one CPCat cassette. There are 1,297 unique CPCat cassettes, including 473 related to drug uses and 824 related to other use categories. The cassettes are permutations of 833 unique CPCat terms, including 456 drug-related terms. Table 2 summarizes the sources with number of original categories, CPCat terms, and chemicals. See http://actor.epa.gov/cpcat for a list of all chemicals included in CPCat and the data dictionary for a list of all CPCat terms and cassettes.

3.2 Example 1: CPCat cassettes associated with a single chemical

The CPCat database can be queried to produce a list of all CPCat terms and cassettes associated with a single chemical. As an example, ethylparaben (Chemical Abstracts Service (CAS) Registry Number (RN) = 120-47-8) is associated with a diverse group of CPCat cassettes (Table 3), most of which are consistent with the use of ethylparaben as a preservative in a variety of cosmetics, soaps, and shampoos. Different cassettes reflect the varying levels of detail present in the original sources categories, which may be important to understand exposure (e.g., "personal_care" vs. "personal_care cosmetics bath baby_use"). Users must also be aware of the "detected" term that may be contained within a CPCat cassette (e.g., "personal_care sexual_wellness gel detected"). This "detected" term indicates the chemical was detected in laboratory tests of the product. Thus, the association of a chemical with a specific product in the database can occur because it is a known ingredient, or because it was detected in laboratory measurements.

343 **Table 3:** CPCat cassettes associated with ethylparaben [a]

CPCat cassettes		
agricultural*	hunting	personal_care cosmetics*
arts_crafts*	industrial cleaning_washing	personal_care sanitizer hand
automotive_care	industrial_manufacturing_ACToRUseDB	personal_care sexual_wellness gel detected
child_use	inert_ACToRUseDB	personal_care shower gel
cleaning_washing*	manufacturing chemical	personal_care soap*
construction	manufacturing cleaning_washing polish	personal_care sunscreen*
consumer_use_ACToRUseDB	manufacturing detergent	personal_care wash*
detergent	manufacturing drug	personal_care_ACToRUseDB
drug*	manufacturing export	pesticide*
electronics batteries*	manufacturing metals	photographic
facility salon detected	manufacturing personal_care*	preservatives
fluid_property_modulator	manufacturing soap	raw_material personal_care cosmetics

food_additive*	paint	sports_equipment
food_additive_ACToRUseDB	paraben	surface_treatment
food_contact	personal_care	tools personal_care hair
fragrance consumer_use	personal_care bath	toys*

[a] A * indicates multiple cassettes containing additional CPCat terms; see Supplemental Material for the full list.

345

346 3.3 Example 2: Child exposure scenario

347 The CPCat database may be queried to identify all chemicals with reported data which fall under
348 a specified exposure scenario. For example, CPCat can be queried for chemicals to which children
349 could be exposed, beyond routine exposures from food, drinking water, dust, and ambient air. To
350 identify such a list of chemicals, we selected CPCat cassettes which include the CPCat terms
351 "baby_use" or "child_use," excluding cassettes including the CPCat terms "food" or
352 "manufacturing" (Table 4). For simplicity cassettes which linked to less than five chemicals were
353 excluded.

354

355 **Table 4:** Selected CPCat cassettes for child exposure scenario

CPCat cassettes	
apparel baby_use diaper	personal_care cosmetics child_use detected
arts_crafts child_use detected	sports_equipment child_use
baby_use detected	toys baby_use
child_use	toys child_use
child_use detected	toys child_use detected
electronics toys child_use	toys fragrance child_use detected
personal_care cosmetics baby_use	toys lawn_garden child_use
personal_care cosmetics bath baby_use	toys mouthing baby_use

356

Extracting the chemicals associated with these 16 cassettes results in 1,074 chemicals mapped to 35 original categories in the RPC, ACToR Data Sets and Lists, and 2012 CDR Consumer database sources. Of these 1,074 chemicals, 649 were associated with the chosen cassettes related to children's exposure based on a single source within CPCat, 211 were associated with the chosen cassettes based on two different sources, and 214 chemicals were associated with the chosen cassettes based on three or more sources. Figure 2 shows a heat map of the CPCat cassettes of interest and associated chemicals for the child scenario. Beyond chemicals associated with the generic "child_use" cassette where no additional descriptors are available, the majority of chemicals in this example are associated with the "toys child_use" CPCat cassette. This indicates that beyond routine exposures from food, drinking water, and ambient air, the largest fraction of chemicals identified were in children's toys. Further, as seen in the gray highlighted bars in the heat map, 386 chemicals are associated with this child exposure scenario through cassettes that include the "detected" CPCat term, but not through any other cassettes. This indicates that if we were researching exposure to chemicals used in children's products, we would be missing potential exposure to 386 chemicals that are not listed as product ingredients but nevertheless were detected in toys and other child-specific products. Information on detection of chemicals in laboratory testing comes from the DCPS source within ACToR Data Sets and Lists, as described above.

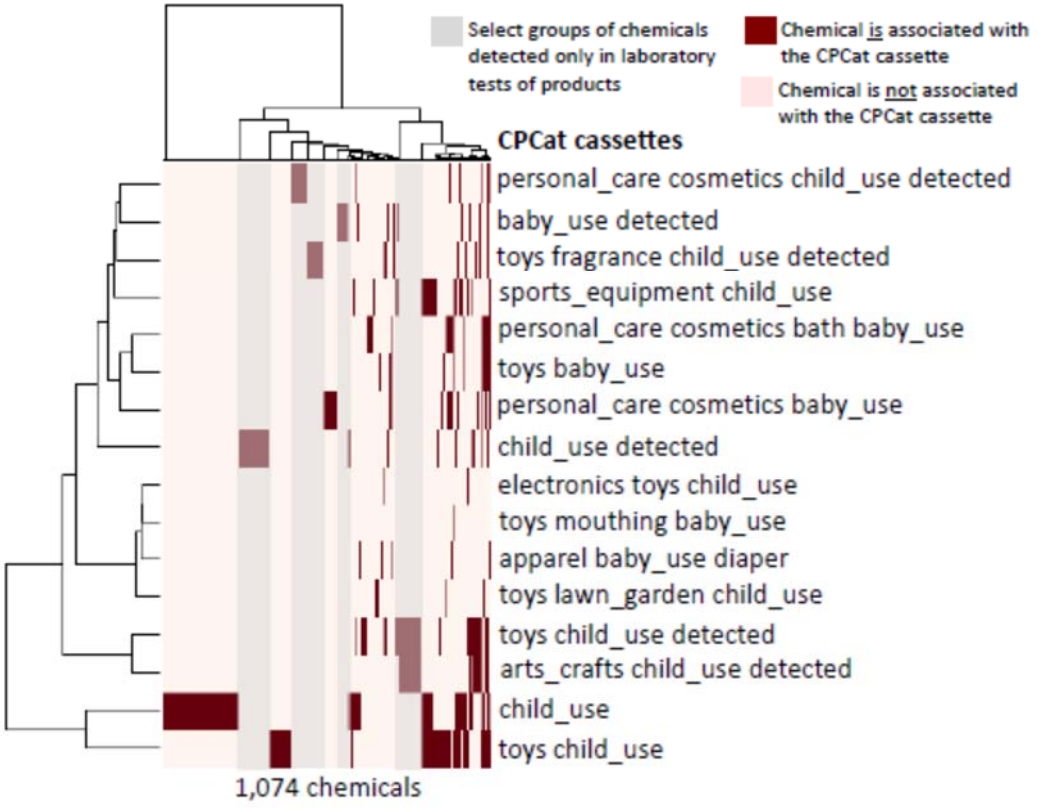


Figure 2: Heat map of chemicals associated with CPCat cassettes from the child scenario. Individual chemicals are on the x-axis, and CPCat cassettes (i.e. use-category classifications) on the y-axis. There are a total of 1,074 chemicals associated with at least one child scenario CPCat cassette.

3.4 Example 3: Potential exposure pathways for chemicals subject to the Endocrine Disruptor Screening Program

CPCat can be queried to identify exposures to chemicals of concern for specific adverse health impacts. For example, the U.S. EPA's Endocrine Disruptor Screening Program (EDSP) is

386 mandated to identify and analyze chemicals for their potential to interact with and disrupt specified
387 endocrine pathways (estrogen, androgen, thyroid and steroidogenesis).
388 The two main classes of chemicals covered in the EDSP are pesticide ingredients (active and inert)
389 and chemicals with the potential to be found in drinking water. This makes up a chemical universe
390 of approximately 5,000 chemicals. In this example, we focus on a set of 5,251 Safe Drinking Water
391 Act (SDWA) chemicals that are candidates for exposure and hazard determination under the EDSP
392 (U.S. EPA 2012).
393 While the CPCat cassettes do not provide any direct, quantitative measure of exposure, they can
394 be used as one input to a prioritization scheme. The first step in exposure prioritization could be
395 to rank the SDWA chemicals by their likely exposure potential, with exposure potential based on
396 the number of consumer-use related CPCat cassettes the chemical is associated with (i.e., the
397 number of consumer-use related "hits"). Theoretically, the more consumer-use related CPCat
398 cassettes that a chemical is associated with would translate to a larger number of potential exposure
399 pathways for an individual (U.S. Department of Health and Human Services 2005). Of all unique
400 CPCat cassettes, 234 were selected as being broadly related to consumer exposure (including
401 exposures from food; Table 5). These 234 consumer exposure related CPCat cassettes are
402 associated with 19,552 unique chemicals.
403
404

405 **Table 5:** Consumer-use related CPCat cassettes selected for EDSP example [a]

CPCat cassettes		
adhesive consumer_use*	drinking_water_contaminant*	lubricant consumer_use*
air_fresheners consumer_use*	electronics*	personal_care*
air_treatment consumer_use	explosives consumer_use	personal_care ACToRUseDB
apparel*	extermination consumer_use	pesticide consumer_use
apparel_care*	fertilizer consumer_use	pet
appliance consumer_use*	flame_retardant	polish apparel_care footwear
arts_crafts*	food*	solvent consumer_use
automotive_care consumer_use	food_additive*	sports_equipment*
automotive_component consumer_use*	food_contact*	stoves consumer_use
baby_use detected*	food_residue*	surface_treatment consumer_use
batteries consumer_use	fragrance consumer_use	tea_coffee
beverage*	fuel automotive	textile consumer_use*

building_material consumer_use*	fuel consumer_use	toilets baby_use
child_use*	fungicide consumer_use	tools consumer_use*
cleaning_washing*	furniture*	tools lawn_garden
colorant consumer_use detected	heating*	tools personal_care*
consumer_use	hunting	toys*
consumer_use_ACToRUseDB	impregnation consumer_use detected	water_treatment consumer_use
décor*	lawn_garden consumer_use	writing*
drinking_water*	leather consumer_use	

406 [a] A * indicates multiple cassettes containing additional CPCat terms. See Supplemental Material for the full list.

Of the 5,251 SDWA compounds, CPCat contains data on 4,189, and 3,514 map to at least one of the consumer-use related CPCat cassettes. Table 6 provides the number of different consumer-use related CPCat cassettes for each of the 22 SDWA chemicals with ≥60 hits. These chemicals could be placed higher on the priority list based on exposure potential, while those compounds which are associated with <5 CPCat cassette hits (2,441 compounds) could be given a lower priority for assessment. It is important to again note that the number of "hits" should not be taken as a quantitative surrogate for exposure measurement, this data can be useful in prioritizing chemicals of interest. A larger number of hits (e.g. ≥ 60 hits versus chemicals with < 5 hits) translates to more confidence in the strength of the evidence that the chemical is included in a variety of consumer-use related products. If we do not have the ability to discriminate between consumer products with high or low exposure dose potential, the presence of the chemical in a large number of products may be a plausible surrogate for an increased probability of exposure. In addition, hits on specific groups of CPCat cassettes could be prioritized based on their exposure potential. For example, if chemicals with fewer hits are included in cassettes with a high exposure potential (e.g., food related CPCat cassettes), those chemicals could be prioritized over chemicals with more hits on cassettes with a lower exposure potential (e.g., cassettes related to apparel).

428 **Table 6**: Number of consumer-use related CPCat cassettes that EDSP/SDWA chemicals are
429 associated with. Chemicals associated with less than 60 consumer related CPCat cassettes are
430 omitted.

CAS RN	Name	CPCat cassette hits
57-55-6	1,2-propanediol	121
64-17-5	ethanol	114
56-81-5	glycerol	110
67-63-0	isopropyl alcohol	90
77-92-9	citric acid	85
99-76-3	methyl 4-hydroxybenzoate	85
1310-73-2	sodium hydroxide	84
13463-67-7	titanium dioxide	82
7647-14-5	sodium chloride	80
102-71-6	2,2,2-nitrilotriethanol	78
106-97-8	butane	74
75-28-5	isobutane	73
94-13-3	propyl 4-hydroxybenzoate	72
128-37-0	2,6-di-tert-butyl-p-cresol	72
3844-45-9	dihydrogen (ethyl)[4-[4-[ethyl(3-sulphonatobenzyl)]amino]-2-sulphonatobenzhydrylidene]cyclohexa-2,5-dien-1-ylidene](3-	65

	sulphonatobenzyl)ammonium, disodium salt	
122-99-6	ethylene glycol monophenyl ether	64
1934-21-0	trisodium 5-hydroxy-1-(4-sulphophenyl)-4-(4-sulphophenylazo)pyrazole-3-carboxylate	64
67-64-1	acetone	63
2682-20-4	2-methyl-4-isothiazolin-3-one	63
14807-96-6	talc ($Mg_3H_2(SiO_3)_4$)	63
100-51-6	benzyl alcohol	62
57-11-4	stearic acid	60

We can further reduce the list of chemicals with a high exposure potential in Table 6 by eliminating chemicals that are common food substances (e.g., ethanol, sodium chloride, citric acid) or are otherwise widely used and considered safe (e.g., talc or other substances on the U.S. FDA's Generally Recognized as Safe (GRAS) list). However, we also see that prioritizing based on the number of consumer-use related CPCat cassette hits does highlight certain phenol compounds that, in their parent or metabolite form, may interact with the estrogen receptor (e.g., propyl 4-hydroxybenzoate, methyl 4-hydroxybenzoate).

4. Discussion

Here we have detailed the construction of the CPCat database, and provided examples of its utility for understanding potential sources of exposure for chemicals in the environment. CPCat contains use information (general-use, product-use, functional-use, therapeutic-use, industrial sector-use) on over 43,000 chemicals taken from major national and international data sources. Of particular note, we have identified a total of ~20,000 unique chemicals with consumer uses. CPCat provides information that one could use to prioritize further study of these chemicals for exposure potential.

There are a number of limitations of the CPCat database that should be taken into account with any use. First, though data from sources such as DrugBank, RPC, and DCPS were hand curated by their respective sources, as described in Methods, there was limited manual curation of data done by the authors, and detailed information about categorizations taken from the original sources was not always available. Besides ACToR, the largest contributor to CPCat is SPIN, and the origin of the data, including how it was identified and collected, is not always clear. Even with ACToR, we have taken data from a large number of other smaller sources, again with limited manual curation. Therefore, it is best to take into account data quality and provenance as appropriate for a particular use, as errors and omissions in the original sources are carried forward in CPCat. However, by including multiple sources of information, one can gain confidence in a general category assignment, especially if the same use category arises from multiple sources.

Another limitation of the CPCat database is that certain category assignments may not equate with bioavailability or potential for exposure. For example, chemicals in CPCat may be assigned to a fabric dye related cassette. It may be assumed by an investigator that individuals may have dermal exposure to these chemicals through clothing that is in direct contact with their skin;

however these chemicals may be tightly bound to the fabric, and thus are likely not bioavailable. Nonetheless, being able to enumerate "all" potential use associations of a chemical has intrinsic value in prioritizing research geared toward elucidating relevant exposure routes, exposure points and exposure pathways from source to receptor.

Lastly, it is important to remember that the CPCat database contains only partial information on the quantities of chemicals in products (namely all information from the RPC Database (Goldsmith et al. 2014)). As shown in Wambaugh et al., the presence or absence of a chemical in consumer products is often an indicator of detection of the chemical in biomonitoring of humans, however uses should recognize that the presence of a chemical mapped to CPCat cassettes is a necessary step in identifying potential exposures, but it is likely insufficient for quantifying exposure.

We envision that a main use of CPCat will be for priority setting tasks, such as in Example 3. The CPCat database can be used to group chemicals by potential types of exposure sources (e.g., by selecting chemicals associated with consumer-use related CPCat cassettes), or by a large number of diverse potential sources (e.g., chemicals associated with a large number of unrelated CPCat cassettes). While CPCat cassettes and terms should not be used as a surrogate for exposure on their own, they can provide information that will aid investigators in identifying chemicals of interest for more detailed analysis. CPCat may also provide intrinsic value in other efforts including systems analysis of key input-output variables in exposure pathways, and ultimately life cycle assessment (LCA) analysis. In the case of LCA, CPCat could greatly assist with identifying inventory flows and processes from the technosphere (man-made world) and systems boundaries used in LCA in a logical, chemical-centric workflow (U.S. EPA 2006).

An interesting potential use of this data comes in exposure modeling. There are existing exposure models that determine population level exposure to a chemical by aggregating doses

from years of simulated individual human interactions with a variety of exposure pathways as they navigate the activities of daily life. These simulations require that there is sufficient data to determine what aspects of daily life may lead to exposure to a specific chemical (Price and Chaisson 2005). For instance, one could add into such a model the sets of chemicals in the "child" scenario described above. Although there are a large number of chemicals, many of them are functional equivalents, so a given person would likely be exposed to one in the class, but not all. In modeling and simulation uses such as these, it will be important to define the functionally equivalent chemicals in a scenario, and perhaps run multiple simulations with different selections out of the equivalent sets.

5. Conclusions

In the absence of more detailed quantitative data on product composition, relevant dose from product use, and exposure routes, CPCat represents a major step forward in characterizing human exposure by making available chemical-to-product use category information. We plan for CPCat to be a continually expanding resource for exposure research. Plans for future work include developing an ontology of exposure and relating the CPCat terms/cassettes to a set of delivery modes (e.g., exposure from a cleaning spray may come through dermal contact with the mixture during cleaning, subsequent ingestion from hand-to-mouth contact, or from inhalation when the mixture is sprayed on a surface) and eventually to exposure models. Other sources of exposure information such as further chemical-to-product mappings (i.e., linkages between retail products and the chemicals contained within the products) can be included, which would enhance the utility of the CPCat database by including quantitative information on the amount of chemicals included in various products. The CPCat database is easily extended, by adding new data, categories or cassettes. Other users could develop and implement their own set of terms or

cassettes, which could be integrated into the current CPCat. We believe that this publicly available database will be a valuable resource for regulators, risk assessors and exposure scientists with a need to evaluate the safety of chemicals.

6. References

Ankley G, Bennett R, Erickson R, Hoff D, Hornung M, Johnson R, et al. 2010. Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. Environ Toxicol Chem 29:730-741.

Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. 2007. The toxcast program for prioritizing toxicity testing of environmental chemicals. Toxicological Sciences 95:5-12.

Egeghy PP, Judson R, Gangwal S, Mosher S, Smith D, Vail J, et al. 2012. The exposure data landscape for manufactured chemicals. Science of the Total Environment 414:159-166.

Goldsmith M-R, Grulke CM, Brooks RD, Transue TR, Tan YM, Frame A, et al. 2014. Development of a consumer product ingredient database for chemical exposure screening and prioritization. Food and Chemical Toxicology 65:269-279.

Judson R, Richard A, Dix D, Houck K, Elloumi F, Martin M, et al. 2008. Actor -- aggregated computational toxicology resource. Toxicology and Applied Pharmacology 233:7-13.

Judson R, Richard A, Dix DJ, Houck K, Martin M, Kavlock R, et al. 2009. The toxicity data landscape for environmental chemicals. Environmental Health Perspectives 117:685-695.

Judson RS, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Mortensen HM, et al. 2010. *In vitro* screening of environmental chemicals for targeted testing prioritization: The toxcast project. Environmental Health Perspectives 118:485-492.

Judson RS, Martin MT, Egeghy P, Gangwal S, Reif DM, Kothiya P, et al. 2012. Aggregating data for computational toxicology applications: The u.S. Environmental protection agency (epa) aggregated computational toxicology resource (actor) system. International Journal of Molecular Sciences 13:1805-1831.

Kavlock R, Dix D. 2010. Computational toxicology as implemented by the u.S. Epa: Providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. Journal of Toxicology and Environmental Health, Part B: Critical Reviews 13:197-217.

Kavlock R, Chandler K, Houck K, Hunter S, Judson R, Kleinstreuer N, et al. 2012. Update on epa's toxcast program: Providing high throughput decision support tools for chemical risk management. Chem Res Toxicol 25:1287-1302.

Mitchell J, Arnot JA, Jolliet O, Georgopoulos PG, Isukapalli S, Dasgupta S, et al. 2013. Comparison of modeling approaches to prioritize chemicals based on estimates of exposure and exposure potential. Science of the Total Environment 458-460:555-567.

Price PS, Chaisson CF. 2005. A conceptual framework for modeling aggregate and cumulative exposures to chemicals. Journal of Exposure Analysis and Environmental Epidemiology 15:473-481.

SPIN. 2013. Spin: Substances in preparations in nordic countries. Available: http://www.spin2000.net [accessed July 9, 2013.

Tice RR, Austin CP, Kavlock RJ, Bucher JR. 2013. Improving the human hazard characterization of chemicals: A tox21 update. Environmental Health Perspectives 121:756-765.

U.S. Department of Health and Human Services. 2005. Public health assessment: Guidance manual (update). Atlanta, Georgia:U.S. Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry.

U.S. EPA. 2006. Life cycle assessment: Principles and practice. Cincinnati, Ohio:National Risk Management Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency.

U.S. EPA. 2012. U.S. Environmental protection agency endocrine disruptor screening program: Universe of chemicals and general validation principles. Washington, D.C.

Wambaugh JF, Setzer RW, Reif DM, Gangwal S, Mitchell-Blackwood J, Arnot JA, et al. 2013. High-throughput models for exposure-based chemical prioritization in the expocast project. Environmental Science & Technology 47:8479-8488.

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. 2006. Drugbank: A comprehensive resource for *in silico* drug discovery and exploration. Nucleic Acids Research 34:D668-D672.

Wishart DS. 2008. Drugbank and its relevance to pharmacogenomics. Pharmacogenomics 9:1166-1162.

Wishart DS, Knox C, guo AC, Cheng D, Shrivastava S, Tzur D, et al. 2008. Drugbank: A knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Research 36:D901-D906.